IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Noam Gal
03/06/2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
  - The data was collected from a SpaceX API, then wrangled using pandas and MySQL
  - Further analysis was conducted with geospatial library folium and machine learning library sklearn
- Summary of all results
  - Built several ML models that showed a 0.83 accuracy score in classifying launches as successes vs. failure
  - In hindsight, ML models do have some predictive power on classifying launches as successes or failures, but that may be due to data on launch sites that are not available in advance

Introduction

- Project background and context
  - This presentation is a capstone project for an IBM Data Science Professional Coursera certificate
- Problems you want to find answers
  - How accurately can ML classifiers predict whether SpaceX Falcoln-9 launches will succeed or fail?
  - What data points are most predictive of a given launch's success?

Section 1

Methodology

Methodology

Executive Summary

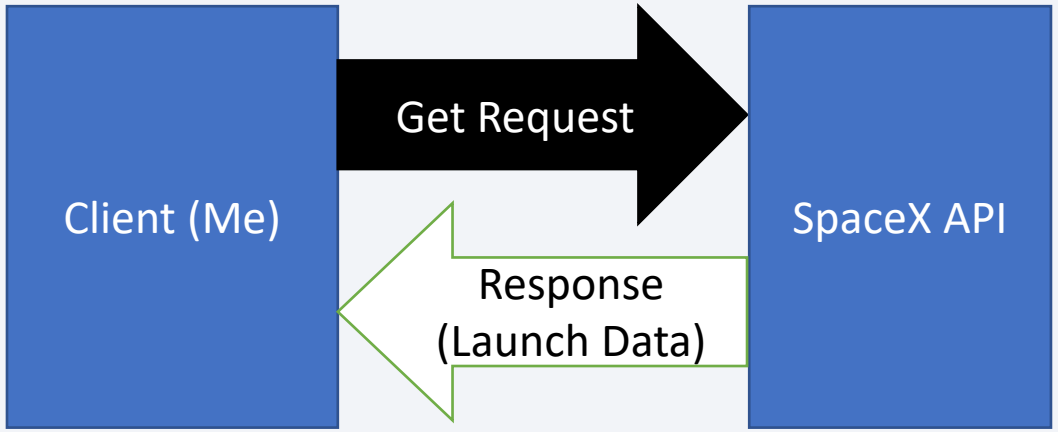
- Data collection methodology:
  - The launch data was collected through a get request to the SpaceX API and by webscraping a Wikipedia page that lists Falcoln rocket launches
- Performed data wrangling
  - The launch data was filtered to include only Falcoln-9 launches and prepared for predictions on launch outcomes
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - After preparing the data, classification models were built using the following algorithms: K-Nearest Neighbors, decision tree, logistic regression and Support Vector Machine

Data Collection

- The SpaceX dataset was retrieved from a REST API in JSON format
  - After parsing the file into a dataframe, I performed basic data wrangling including replacing null values and filtering the data to only Falcoln-9 launches
- A secondary dataset was webscraped in HTML format from a Wikipedia page listing Falcoln rocket launches
  - I used the BeautifulSoup library to parse the html file into a dataframe

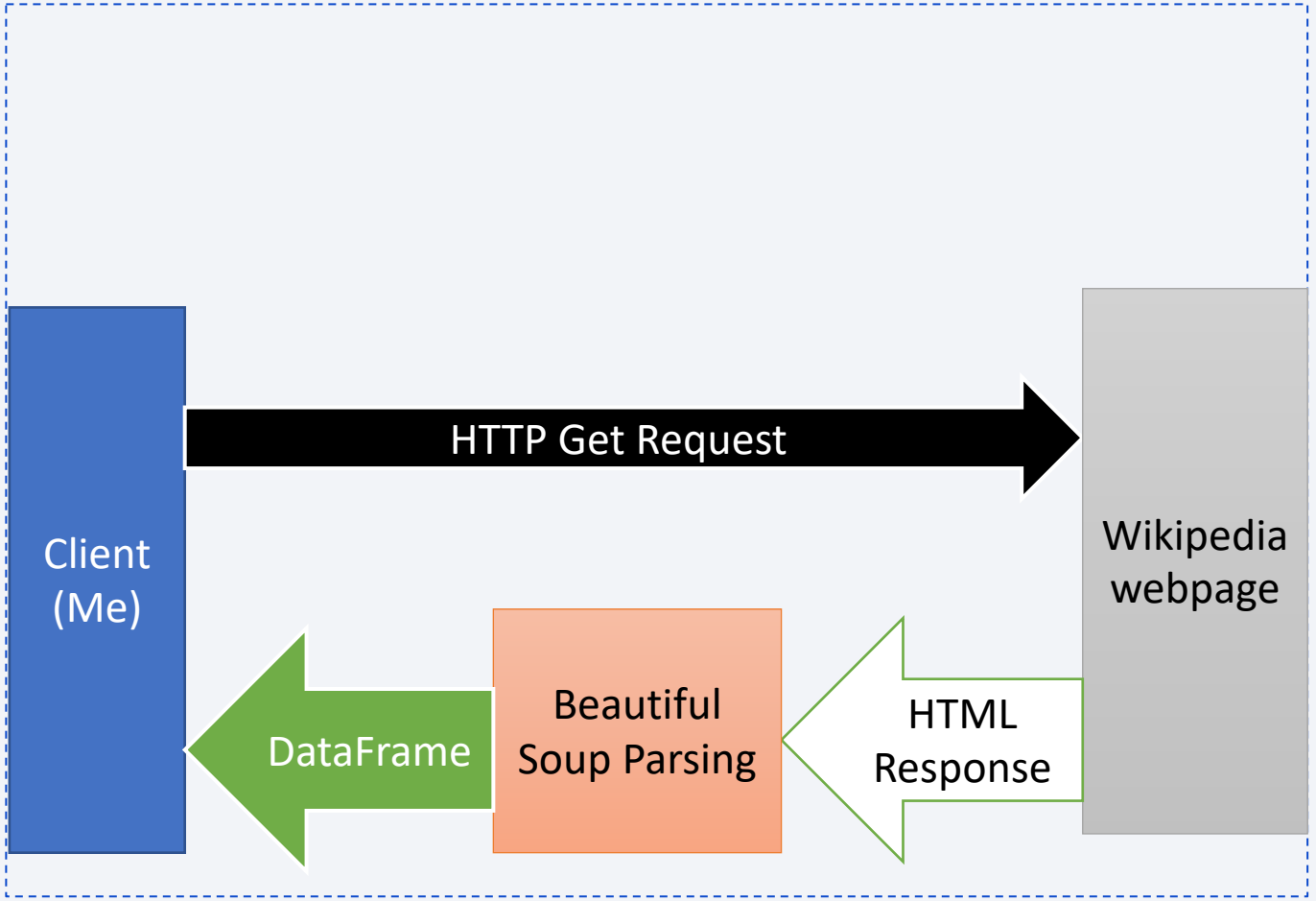
Data Collection – SpaceX API

- Used a get request to the SpaceX API using the following url https://api.spacexdata.com/v4/launches/past which returned launch data on SpaceX's Falcoln-1 and Falcoln-9 rockets
- **Reference Notebook:** https://github.com/noamjgal/Data-Science-Pro-IBM/blob/d8da007f304963d991e586006722ac70cbefd8b4/SpaceX-Final/1-Data-Collection-SpaceX-API.ipynb



Data Collection - Scraping

- Used a get request to retrieve the HTML version of a Wikipedia page that details Falcoln 9 launches and
- Parsed the HTML file using the BeautifulSoup library to create a dataframe where each row corresponds to details on a different launch
- **Reference Notebook:** https://github.com/noamjgal/Data-Science-Pro-IBM/blob/d8da007f304963d991e586006722ac70cbefd8b4/SpaceX-Final/2-Falcoln9-Webscraping.ipynb



Data Wrangling

- Prepared the data to train the binary classification algorithms by categorizing all launches as successes or failures based on the outcomes of the landing
- https://github.com/noamjgal/Data-Science-Pro-IBM/blob/d8da007f304963d991e586006722ac70cbefd8b4/SpaceX-Final/3-Landing-Data-Wrangling.ipynb

EDA with Data Visualization

- Created a series of scatter plots that demonstrate the relationships between different factors in each launch including payload mass, target orbits, flight number, and launch site
- Created a bar plot showing the success rates of each orbit type
- Created a line plot visualizing how the success rate of Falcoln9 launches has changed over time
- https://github.com/noamjgal/Data-Science-Pro-IBM/blob/407f4cafe389409ee2922dd2c98650dd69cf9538/SpaceX-Final/4-Falcoln9-EDA-Visualization.ipynb

EDA with SQL

- Used an SQL extension called sql alchemy to query SpaceX launch data stored in the IBM database
- Computed descriptive data about the launches such as total payload mass, average payload mass, dates and booster versions for different launches, etc.
- GitHub URL: https://github.com/noamjgal/Data-Science-Pro-IBM/blob/7d596e44dcf6005cc812826c13625a438d9b591c/SpaceX-Final/5-SQL-Landing-Outcome-Analysis.ipynb

Build an Interactive Map with Folium

- Mapped the launch sites with clusters of markers showing each launch in green if successful and in red if unsuccessful
- Added lines showing distances to a nearby railroad, high-way, and city for the Cape Canaveral launch site for added detail on the surrounding area
- https://github.com/noamjgal/Data-Science-Pro-IBM/blob/main/SpaceX-Final/6-Folium-Launch-Site-Mapping.ipynb
- https://github.com/noamjgal/Data-Science-Pro-IBM/blob/main/SpaceX-Final/6-Folium-Launch-Site-Mapping.pdf

Build a Dashboard with Plotly Dash

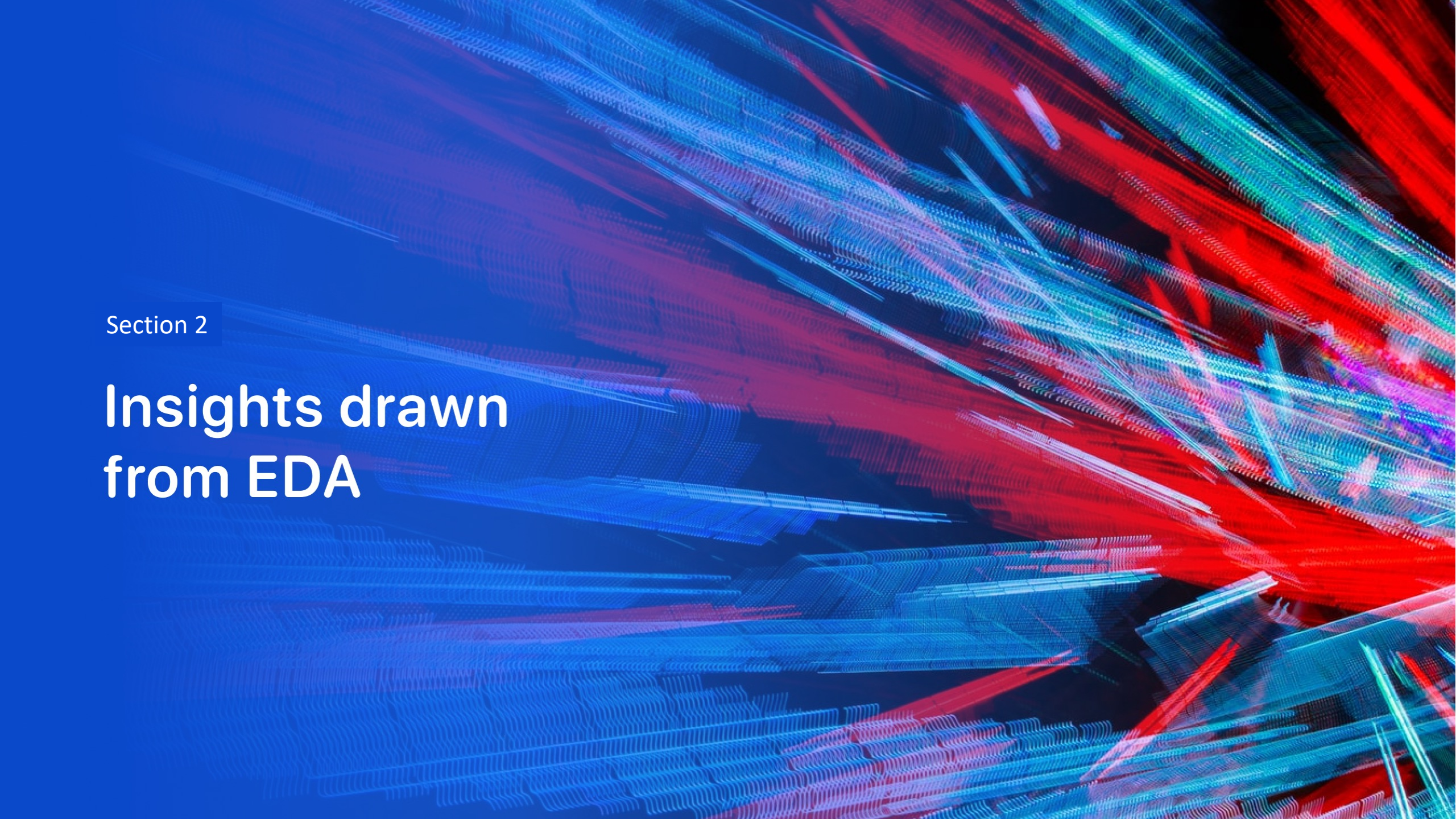
- Created an interactive pie chart to visualize both the distribution of launch successes between sites and the success rate at each site
- Created an interactive scatter plot that provides the payload mass, success/failure label, and launch site for every launch
- https://github.com/noamjgal/Data-Science-Pro-IBM/blob/62e080afdc45ed1921b7f77af9410261be2a2581/SpaceX-Final/7-SpaceX-Dash-Plot.py
- https://github.com/noamjgal/Data-Science-Pro-IBM/blob/62e080afdc45ed1921b7f77af9410261be2a2581/SpaceX-Final/7-SpaceX-Dash-Plot-Printout.pdf

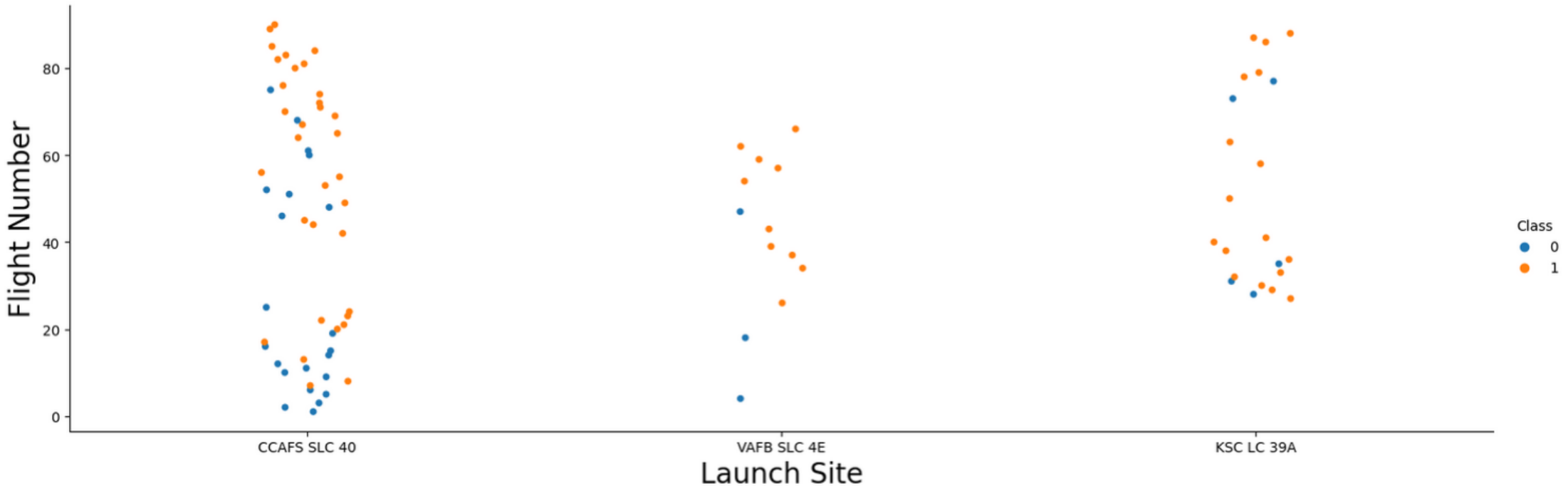
Predictive Analysis (Classification)

- Built, optimized and trained four classification algorithms on the launch data to output a binary label of success or failure for each model using the sklearn library
- Split the data into training and testing sets, then used cross validated grid search on the training data to located the optimal hyperparameters for each algorithm
- All four algorithms operated equivalently well on test data with an accuracy score of ~0.83
- Plotted the results of the project with confusion matrices and a bar chart of the algorithms' accuracy scores
- https://github.com/noamjgal/Data-Science-Pro-IBM/blob/62e080afdc45ed1921b7f77af9410261be2a2581/SpaceX-Final/8-SpaceX-ML-Prediction.ipynb

Results

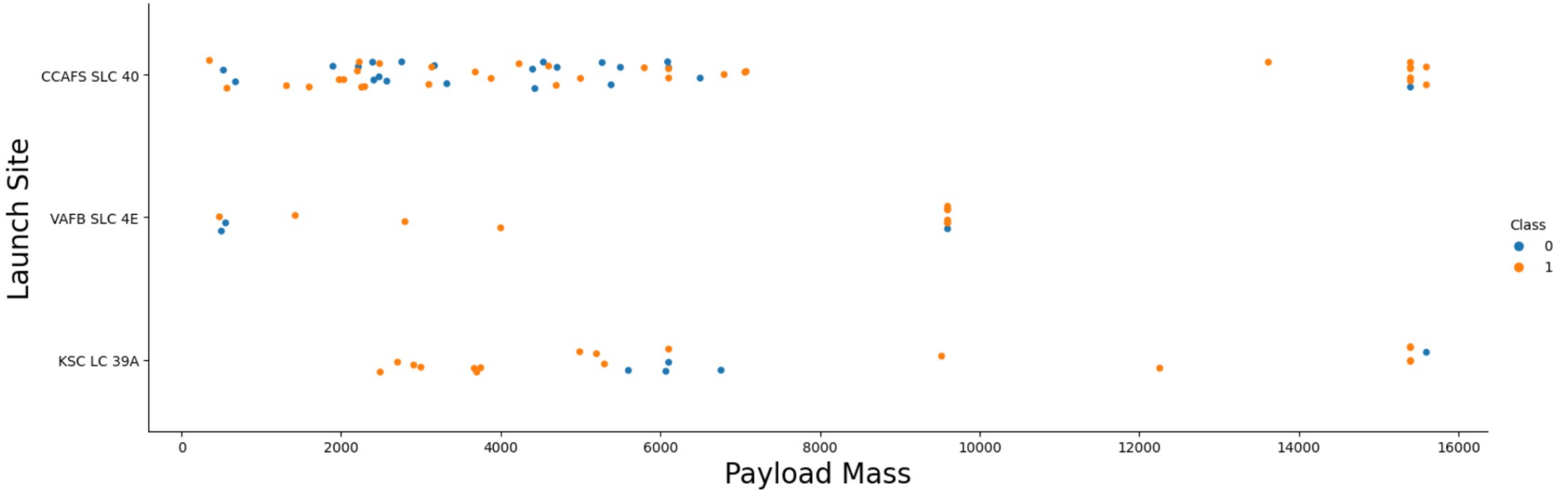
- Exploring the dataset, I found that the Falcoln-9 project appears to be improving rapidly and is reaching higher target orbits, higher payload masses, and higher success rates over time
- Found that the KSC LC-39A had a significantly higher success rate than any other launch site
- Found that classification algorithms can be effective at predicting whether Falcoln-9 launches will succeed or fail based on descriptive data


Section 2
Insights drawn from EDA



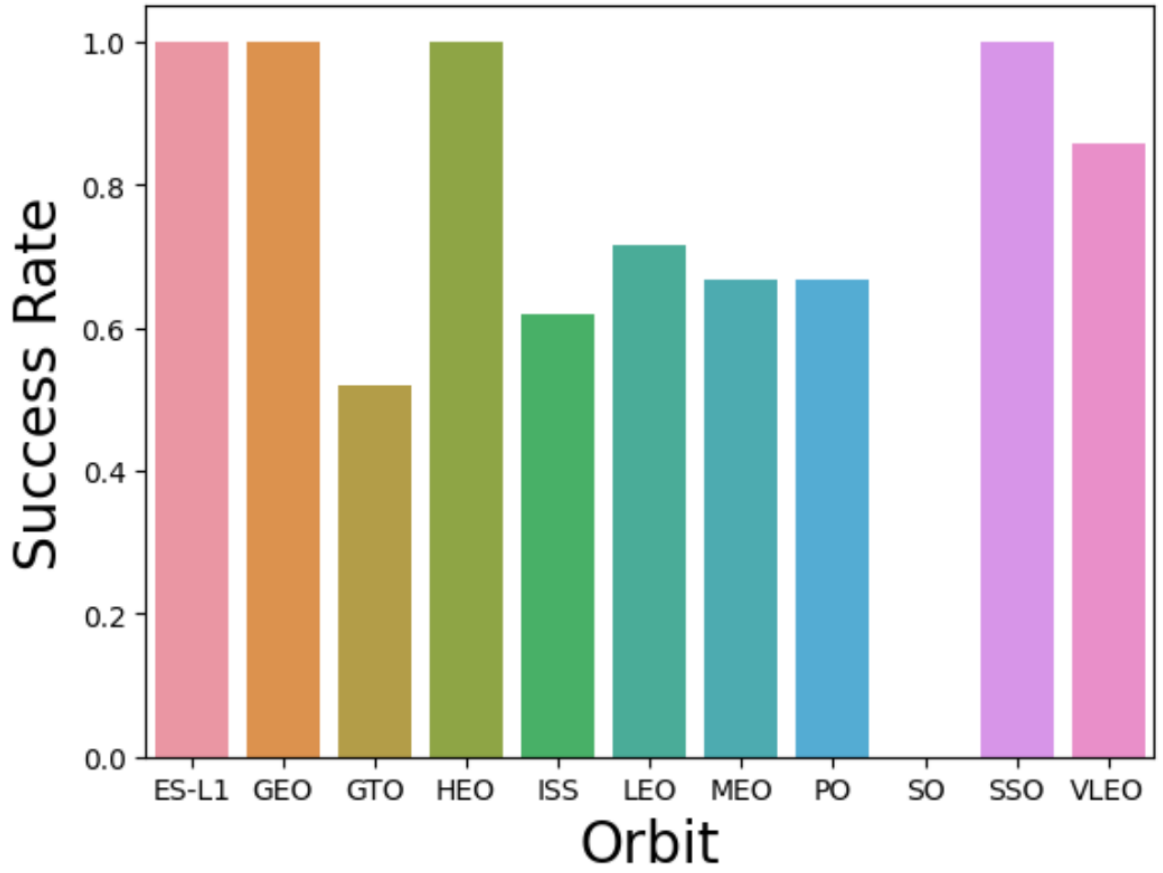
Flight Number vs. Launch Site Scatter Plot

- The flight numbers are relatively evenly distributed at the CCADS SLC 40 site indicating that it operated throughout the project
- The flight numbers tend towards the low end at the VAFB SLC 4E site indicating that the site came online early and was closed
- The flight numbers tend towards the high end at the KSC LC 39A site indicating that the site came online later in the project
- At all three sites, the early launches seem likeliest to fail



Payload vs. Launch Site

- Most launches are under 8,000 kilograms, but there are clusters at the high end of the range at 10000 and 16000
- It is likely that the lower payload launches were tests and were working up to the max payload launches that will be used in standard operations

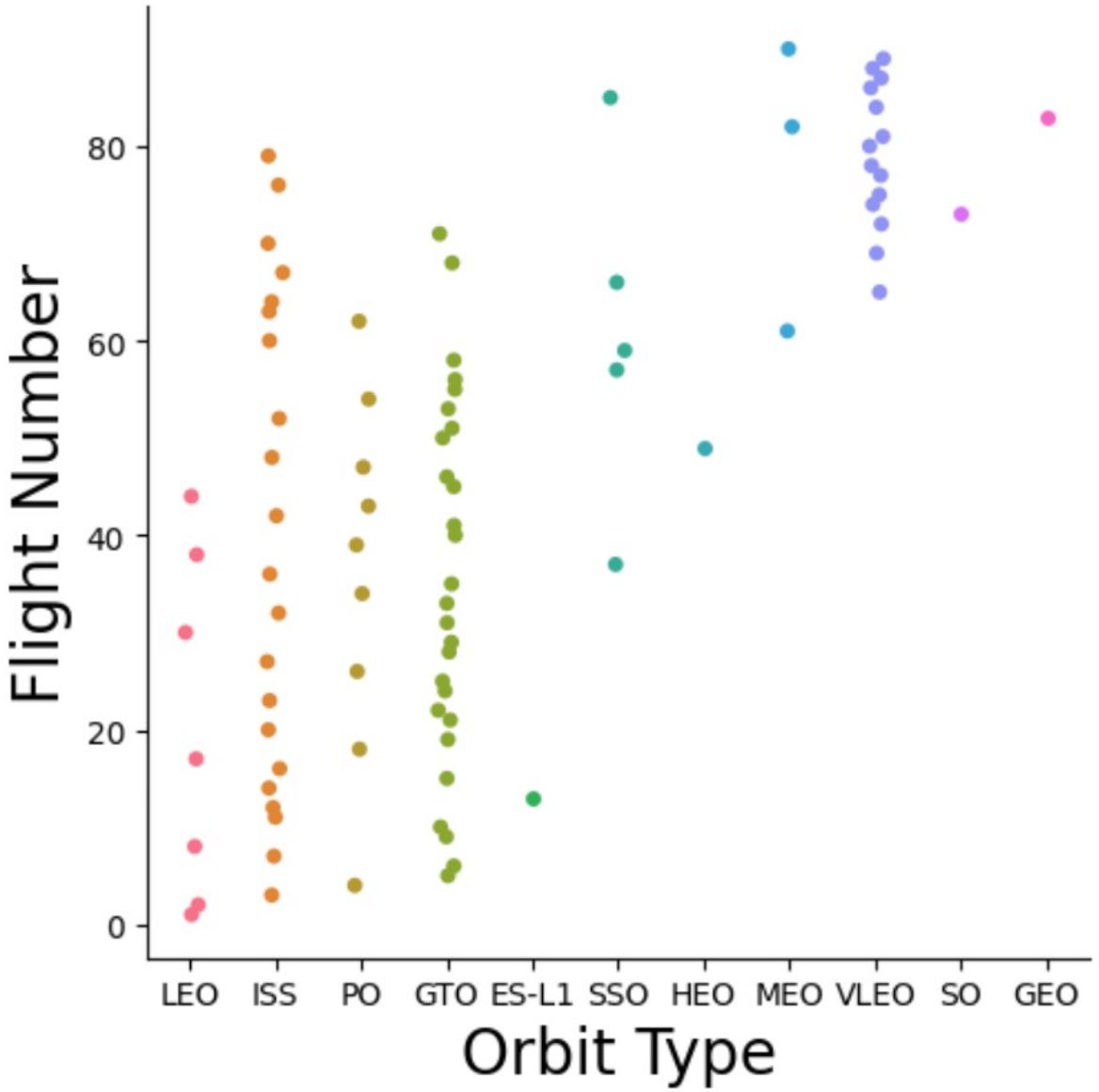


Success Rate vs. Orbit Type

- Certain orbits appear to have far higher success rates than others indicating a strong relationship between orbit type and the likelihood of a launch success

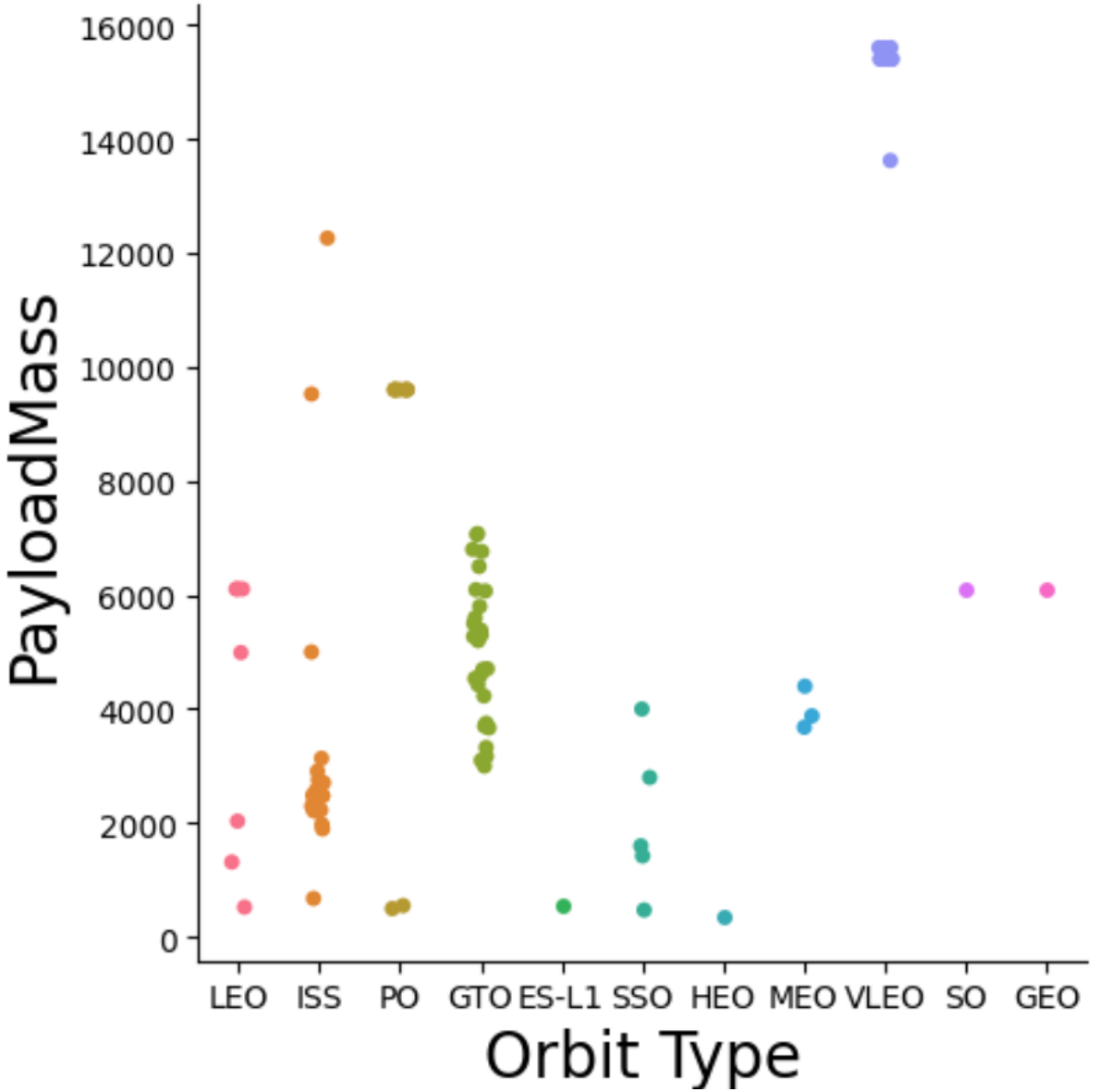
Flight Number vs. Orbit Type

- There is a clear relationship between flight number and orbit types
- Certain orbits are tested on only by the early flights, some orbits are tested on throughout, and some orbits were only tested on by later flights



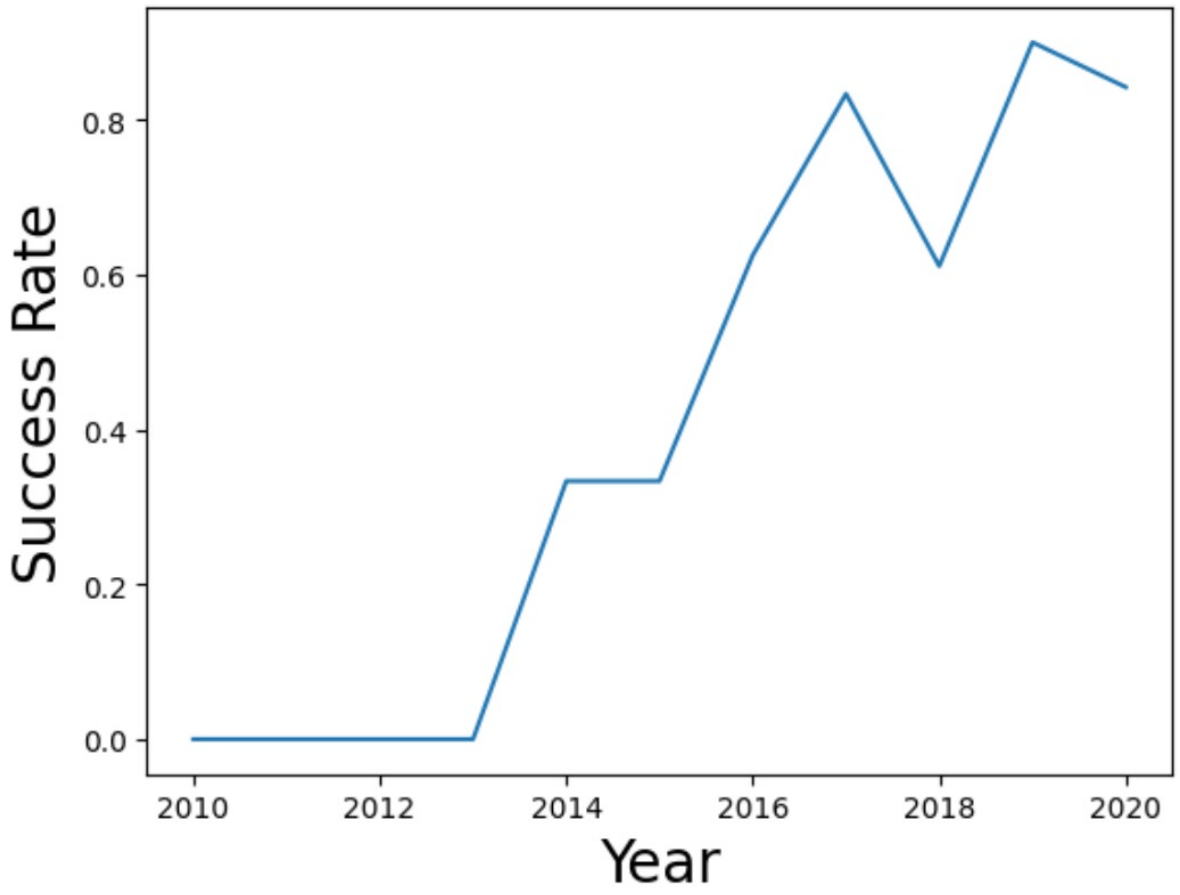
Payload vs. Orbit Type

- Most orbits were only tested on with low payload masses
- The VLEO orbit was tested on only by high payload masses indicating that it is intended as a target orbit for standard operations



Launch Success Yearly Trend

- Launch success has increased over time indicating that the project is succeeding in reaching its goals



SQL Unique Launch Sites

```
In [8]: %sql SELECT DISTINCT Launch_Site from SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

Out[8]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

- The Select Distinct Launch Site query returned the following four sites:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
In [9]: %sql SELECT * from SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

```
 * sqlite:///my_data1.db
Done.
```

Out[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Queried the launch site dataset to filter for 5 site names beginning with CCA

Total Payload Mass

```
In [10]: %sql SELECT sum(payload_mass__kg_) from SPACEXTBL WHERE customer = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
Out[10]:
```

sum(payload_mass__kg_)
45596

- Queried the SpaceX dataset to calculate the total sum of payload masses for NASA launches

Average Payload Mass by F9 v1.1

```
In [11]: %sql SELECT avg(payload_mass__kg_) from SPACEXTBL WHERE booster_version = 'F9 v1.1'
          * sqlite:///my_data1.db
         Done.
Out[11]: avg(payload_mass__kg_)
                       2928.4
```

- Queried the SpaceX dataset to calculate the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
In [12]: %sql SELECT min(DATE) from SPACEXTBL WHERE "Landing _Outcome" = 'Success (ground pad)'
 * sqlite:///my_data1.db
Done.
Out[12]:
```

min(DATE)
01-05-2017

- Queried the SpaceX dataset to find the dates of the first successful landing outcome on ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [18]:
```
%sql SELECT booster_version from SPACEXTBL WHERE "Landing _Outcome" = 'Success (drone ship)' \
and payload_mass__kg_ BETWEEN 4000 and 6000
```

```
 * sqlite:///my_data1.db
Done.
```

Out[18]:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- Queried the SpaceX dataset to return the names of all boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [14]:
```
%sql SELECT mission_outcome, count(mission_outcome) from SPACEXTBL GROUP BY mission_outcome
```

```
 * sqlite:///my_data1.db
Done.
```

Out[14]:

Mission_Outcome	count(mission_outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Queried the SpaceX dataset to calculate the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [15]:
```
%sql SELECT booster_version, payload_mass__kg_ from SPACEXTBL\
WHERE payload_mass__kg_ = (SELECT max(payload_mass__kg_) from SPACEXTBL)
```

```
 * sqlite:///my_data1.db
Done.
```

Out[15]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

Queried the SpaceX dataset to list the names of the booster which have carried the maximum payload mass of 15,600 kilograms

2015 Launch Records

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

In [16]:
```
%sql SELECT substr(Date, 4, 2), "Landing _Outcome", booster_version, launch_site from SPACEXTBL\
WHERE "Landing _Outcome" = 'Failure (drone ship)' and substr(Date,7,4) = '2015'
```

```
 * sqlite:///my_data1.db
Done.
```

Out[16]:

substr(Date, 4, 2)	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Queried the SpaceX dataset to list failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015

Successful Landing Outcomes Between 2010-06-04 and 2017-03-20

- Queried the SpaceX dataset to show all successful landing events and dates between 2010-06-04 and 2017-03-20

Print all successful landing_outcomes with dates

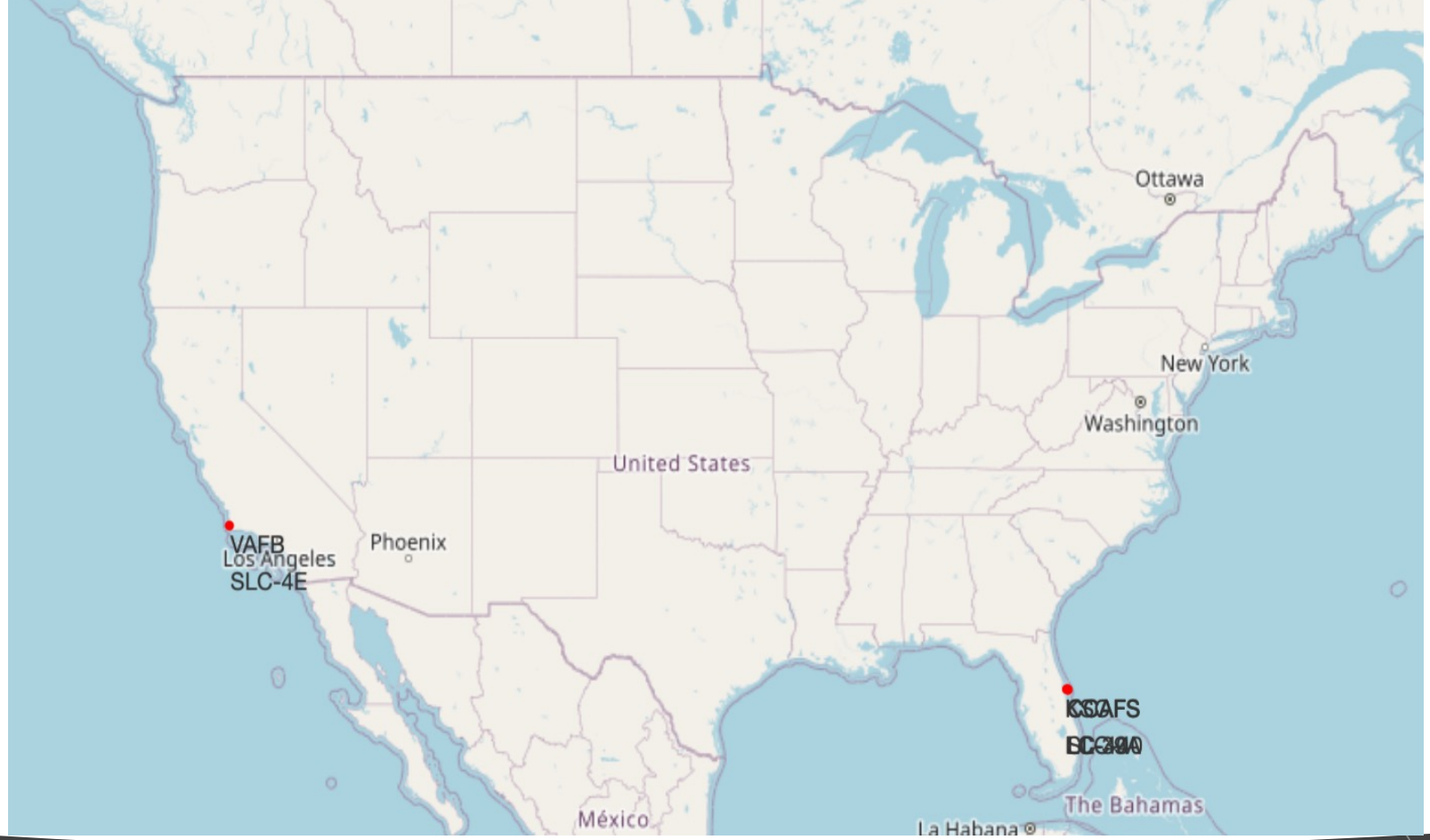
In [24]:
```
%sql SELECT strftime('%m/%d/%Y',datetime(substr(DATE, 7, 4) || '-' || substr(DATE, 4, 2) || '-' || substr(DATE, 1, 2
AS format_date, "Landing _Outcome" from SPACEXTBL \
WHERE  "Landing _Outcome" LIKE "%Success%";
```

```
 * sqlite:///my_data1.db
Done.
```

Out[24]:

format_date	Landing _Outcome
12/22/2015	Success (ground pad)
04/08/2016	Success (drone ship)
05/06/2016	Success (drone ship)
05/27/2016	Success (drone ship)
07/18/2016	Success (ground pad)
08/14/2016	Success (drone ship)
01/14/2017	Success (drone ship)
02/19/2017	Success (ground pad)
03/30/2017	Success (drone ship)
05/01/2017	Success (ground pad)
06/03/2017	Success (ground pad)
06/23/2017	Success (drone ship)
06/25/2017	Success (drone ship)
08/14/2017	Success (ground pad)
08/24/2017	Success (drone ship)
09/07/2017	Success (ground pad)
10/09/2017	Success (drone ship)
10/11/2017	Success (drone ship)
10/30/2017	Success (drone ship)
12/15/2017	Success (ground pad)
01/08/2018	Success (ground pad)
04/18/2018	Success (drone ship)
05/11/2018	Success (drone ship)
07/22/2018	Success
07/25/2018	Success
08/07/2018	Success
09/10/2018	Success
10/08/2018	Success
11/15/2018	Success
12/03/2018	Success
01/11/2019	Success
02/22/2019	Success
03/02/2019	Success
05/04/2019	Success
05/24/2019	Success
06/12/2019	Success
07/25/2019	Success
11/11/2019	Success
12/05/2019	Success
12/17/2019	Success
01/07/2020	Success
01/29/2020	Success
03/07/2020	Success
04/22/2020	Success
05/30/2020	Success
06/04/2020	Success
06/13/2020	Success
06/30/2020	Success
07/20/2020	Success
08/07/2020	Success
08/18/2020	Success
08/30/2020	Success
09/03/2020	Success
10/06/2020	Success
10/18/2020	Success
10/24/2020	Success
11/05/2020	Success
11/16/2020	Success
11/21/2020	Success
11/25/2020	Success
12/06/2020	Success


Section 3
Launch Sites
Proximities Analysis



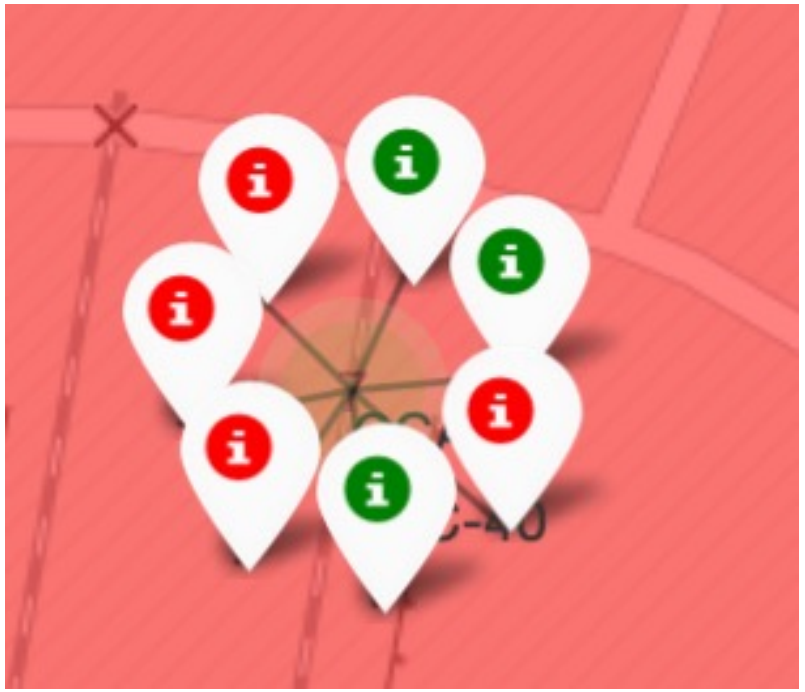
Launch Site Map

- The launch sites can be seen to be concentrated on the coasts of two regions: Southern California and Central Florida

Success Mapping

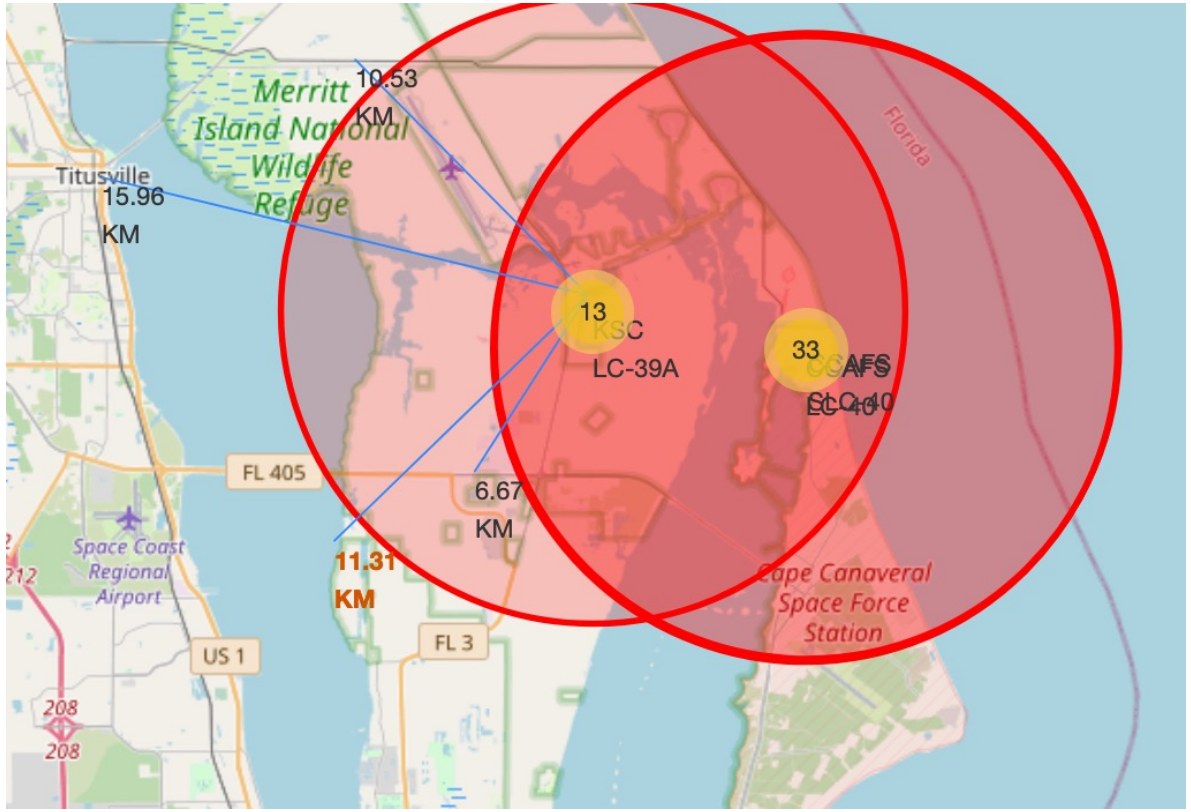
- All four launch site clusters are shown with success labeled in green and failures in red

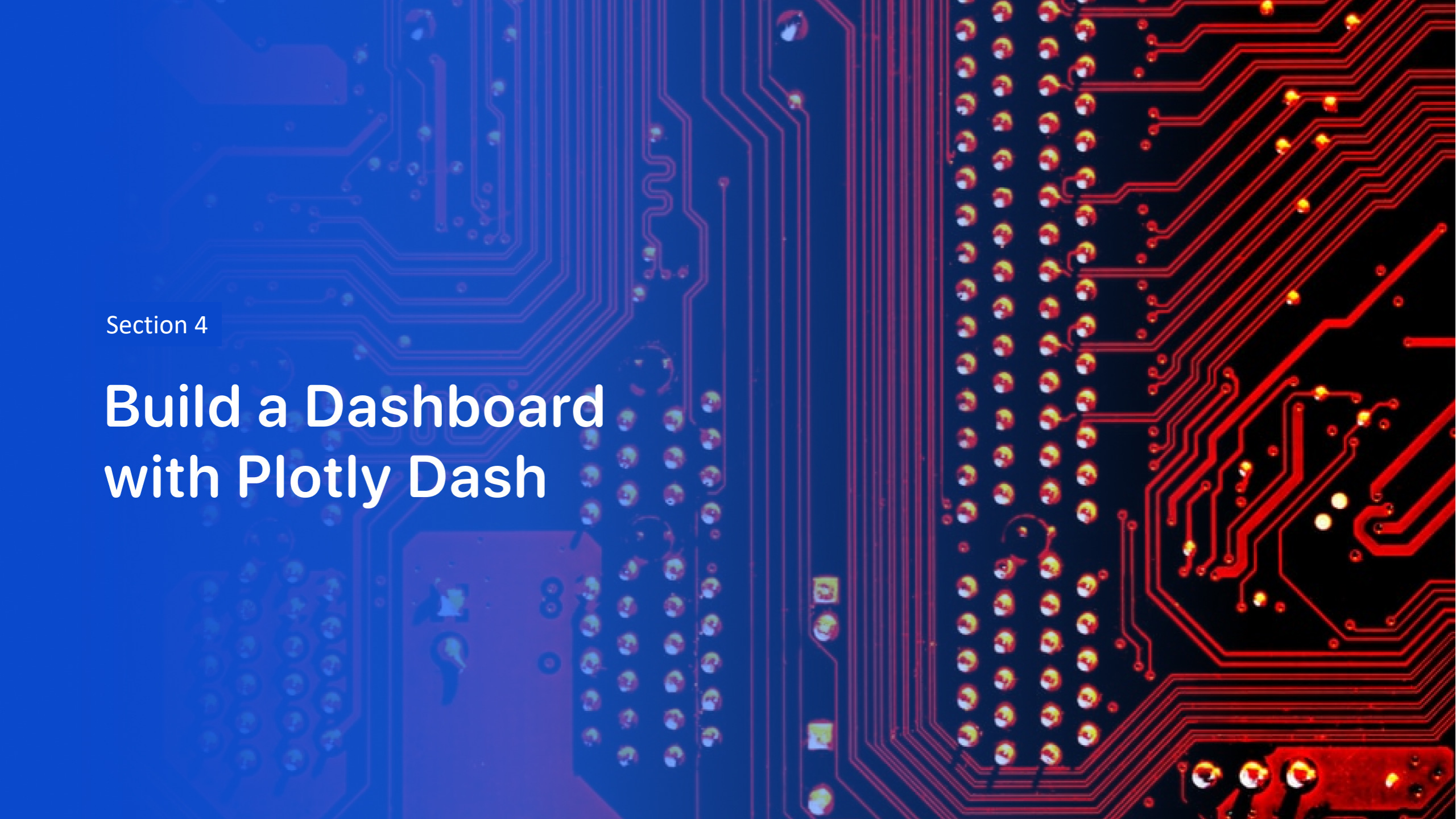
- The KSC LC-39A launch site can be seen to have a far higher success rate than any other



Area Analysis

- Launch sites tend to be close to railways, highways, and coastlines, while maintaining from distance cities.
- Railways and highways are essential for logistics.
- It is important for safety purposes that launches take place near a coastline, to allow for soft landings into the ocean instead of land., and far from cities, since it would be catastrophic if a projectile were to land in an urban area.

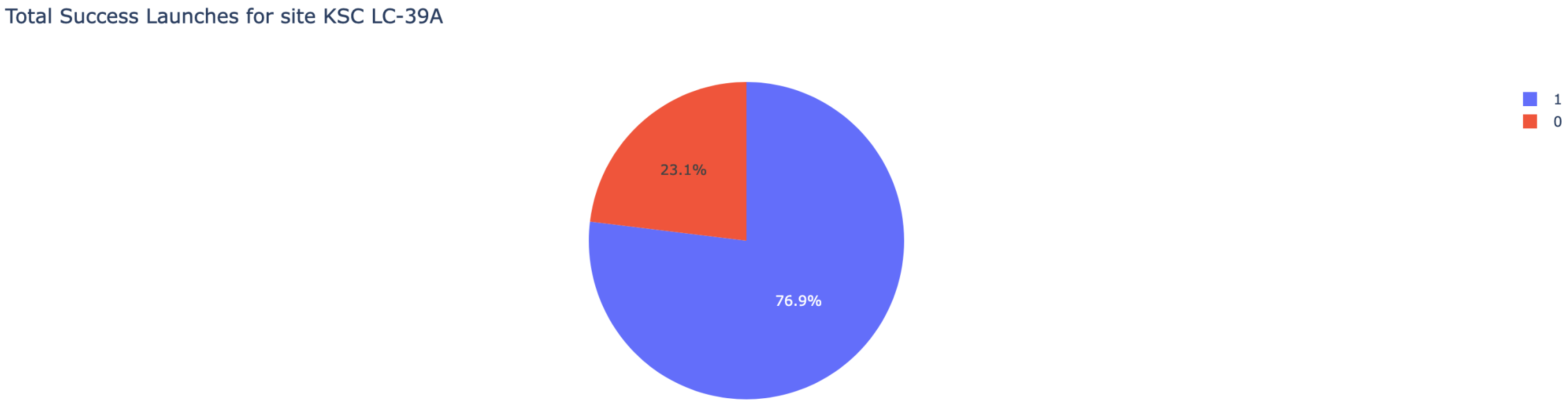



Section 4
Build a Dashboard with Plotly Dash

Success Count for all launch sites

Successful Launch Distribution

- There is an uneven distribution of successful launches between sites, suggesting that specific sites may be more effective
- The sites with more successful launches also have higher success rates (success/total launches)



Success Rate at Top Site

- At the site with the most successful launches, KSC LC-39A, just over three quarters of launches were successful, outperforming any other site

Payload Mass and Launch Success

- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.
- Higher payload masses appeared to have lower success rates

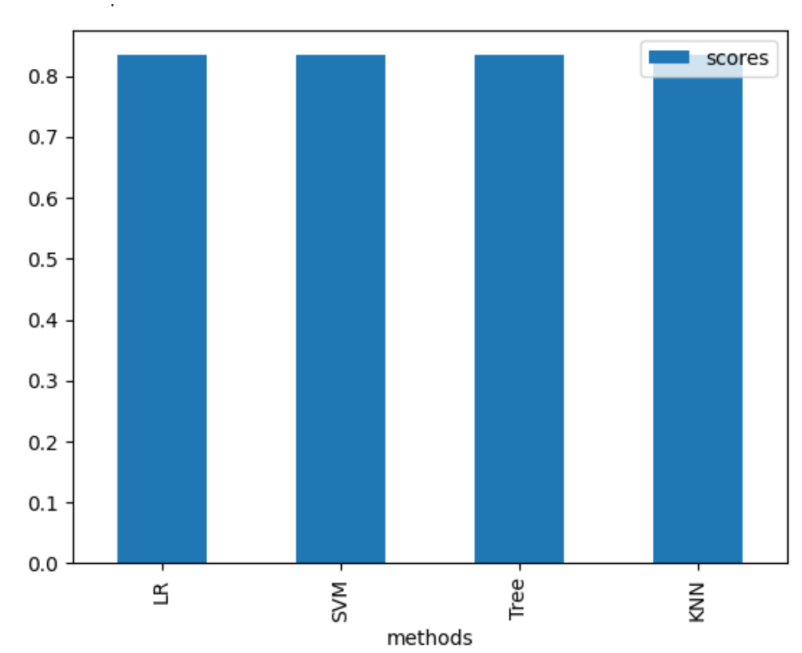
Section 5

Predictive Analysis (Classification)

Classification Accuracy

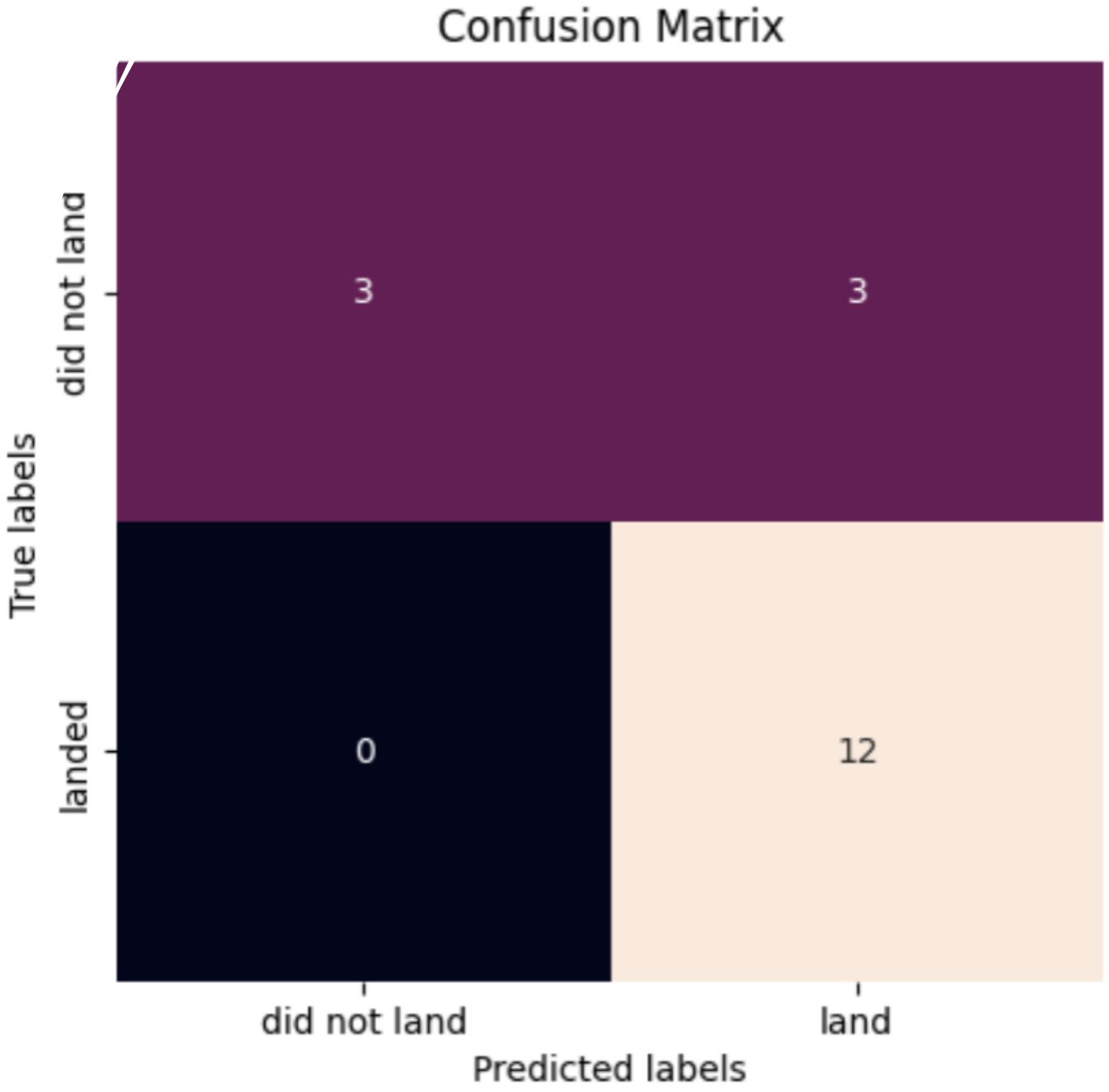
- I found that all four models had the same accuracy score of roughly 0.83 where a perfect accuracy score is 1

methods	scores
LR	0.833333
SVM	0.833333
Tree	0.833333
KNN	0.833333



Confusion Matrix

- All models resulted in the same confusion matrix on the test sample
- The models each predicted three false positives where they labeled a launch as landing when it in fact failed
- There were no false negatives in any model



Conclusions

- Launch success rate has increased significantly over the past ten years indicating that SpaceX's operating procedures and technologies are improving at a rapid clip
- Site locations are an important factor in ensuring the safety of launches and may be a major factor in predicting likelihood of launch success
- The KSC LC-39A launch site in Cape Canaveral has a higher Falcoln-9 launch success rate than any other SpaceX launch site
- All ML algorithms tested on predicting launch success in this project operated at a similar level of effectiveness
- In hindsight, ML models can be effective at classifying launches as successes or failures
- It is possible that the models build in this project are overfitted to the limited available data and further research should be conducted as new data, launch sites, and technologies come online

Thank you!